Neurocomputing 192 (2016) 38–48

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Mean Absolute Percentage Error for regression models

Arnaud de Myttenaere [a,c], Boris Golden [a], Bénédicte Le Grand [b], Fabrice Rossi [c,*]

[a] Viadeo, 30 rue de la Victoire, 75009 Paris - France
[b] Centre de Recherche en Informatique – Université Paris 1 Panthéon - Sorbonne, 90 rue de Tolbiac, 75013 Paris - France
[c] SAMM EA 4534 – Université Paris 1 Panthéon - Sorbonne, 90 rue de Tolbiac, 75013 Paris - France

ARTICLE INFO

Article history:
Received 11 July 2015
Received in revised form
1 November 2015
Accepted 2 December 2015
Available online 10 March 2016

Keywords:
Mean Absolute Percentage Error
Empirical Risk Minimization
Consistency
Optimization
Kernel regression

ABSTRACT

We study in this paper the consequences of using the Mean Absolute Percentage Error (MAPE) as a measure of quality for regression models. We prove the existence of an optimal MAPE model and we show the universal consistency of Empirical Risk Minimization based on the MAPE. We also show that finding the best model under the MAPE is equivalent to doing weighted Mean Absolute Error (MAE) regression, and we apply this weighting strategy to kernel regression. The behavior of the MAPE kernel regression is illustrated on simulated data.

1. Introduction

Classical regression models are obtained by choosing a model that minimizes an empirical estimation of the Mean Square Error (MSE). Other quality measures are used, in general for robustness reasons. This is the case of the Huber loss [1] and of the Mean Absolute Error (MAE, also know as median regression), for instance. Another example of regression quality measure is given by the Mean Absolute Percentage Error (MAPE). If x denotes the vector of explanatory variables (the input to the regression model), y denotes the target variable and g is a regression model, the MAPE of g is obtained by averaging the ratio $\frac{|g(x)-y|}{|y|}$ over the data.

The MAPE is often used in practice because of its very intuitive interpretation in terms of relative error. The use of the MAPE is relevant in finance, for instance, as gains and losses are often measured in relative values. It is also useful to calibrate prices of products, since customers are sometimes more sensitive to relative variations than to absolute variations.

In real world applications, the MAPE is frequently used when the quantity to predict is known to remain way above zero. It was used for instance as the quality measure in a electricity consumption forecasting contest organized by GdF ecometering on datascience.net.[1] More generally, it has been argued that the MAPE is very adapted for forecasting applications, especially in situations where enough data are available, see e.g. [2].

We study in this paper the consequences of using the MAPE as the quality measure for regression models. Section 2 introduces our notations and the general context. It recalls the definition of the MAPE. Section 3 is dedicated to a first important question raised by the use of the MAPE: it is well known that the optimal regression model with respect to the MSE is given by the regression function (i.e., the conditional expectation of the target variable knowing the explanatory variables). Section 3 shows that an optimal model can also be defined for the MAPE. Section 4 studies the consequences of replacing MSE/MAE by the MAPE on capacity measures such as covering numbers and Vapnik–Chervonenkis dimension. We show in particular that MAE based measures can be used to upper bound MAPE ones. Section 5 proves a universal consistency result for Empirical Risk Minimization applied to the MAPE, using results from Section 4. Finally, Section 6 shows how to perform MAPE regression in practice. It adapts quantile kernel regression to the MAPE case and studies the behavior of the obtained model on simulated data.

2. General setting and notations

We use in this paper a standard regression setting in which the data are fully described by a random pair $Z=(X,Y)$ with values in $\mathbb{R}^d \times \mathbb{R}$. We are interested in finding a good model for the pair, that is a (measurable) function g from $\mathbb{R}^d$ to $\mathbb{R}$ such that $g(X)$ is "close to" Y. In the classical regression setting, the closeness of $g(X)$ to Y is

* Corresponding author.
E-mail addresses: ademyttenaere@viadeoteam.com (A. de Myttenaere), bgolden@viadeoteam.com (B. Golden), Benedicte.Le-Grand@univ-paris1.fr (B. Le Grand), Fabrice.Rossi@univ-paris1.fr (F. Rossi).

[1] http://www.datascience.net, see https://www.datascience.net/fr/challenge/16/details for details on this contest.

http://dx.doi.org/10.1016/j.neucom.2015.12.114
0925-2312/

measured via the L_2 risk, also called the mean squared error (MSE), defined by

$$L_2(g) = L_{MSE}(g) = \mathbb{E}(g(X)-Y)^2. \quad (1)$$

In this definition, the expectation is computed by respect to the random pair (X,Y) and might be denoted $\mathbb{E}_{X,Y}(g(X)-Y)^2$ to make this point explicit. To maintain readability, this explicit notation will be used only in ambiguous settings.

Let m denote the regression function of the problem, that is the function from $\mathbb{R}^d$ to $\mathbb{R}$ given by

$$m(x) = \mathbb{E}(Y|X=x). \quad (2)$$

It is well known (see e.g. [3]) that the regression function is the best model in the case of the mean squared error in the sense that $L_2(m)$ minimizes $L_2(g)$ over the set of all measurable functions from $\mathbb{R}^d$ to $\mathbb{R}$.

More generally, the quality of a model is measured via a *loss function*, l, from $\mathbb{R}^2$ to $\mathbb{R}^+$. The point-wise loss of the model g is $l(g(X),Y)$ and the *risk* of the model is

$$L_l(g) = \mathbb{E}(l(g(X),Y)). \quad (3)$$

For example, the squared loss, $l_2 = l_{MSE}$ is defined as $l_2(p,y) = (p-y)^2$. It leads to the L_{MSE} risk defined above as $L_{l_2}(g) = L_{MSE}(g)$.

The *optimal risk* is the infimum of L_l over measurable functions, that is

$$L_l^* = \inf_{g \in \mathcal{M}(\mathbb{R}^d,\mathbb{R})} L_l(g), \quad (4)$$

where $\mathcal{M}(\mathbb{R}^d,\mathbb{R})$ denotes the set of measurable functions from $\mathbb{R}^d$ to $\mathbb{R}$. As recalled above we have

$$L_{MSE}^* = L_2^* = L_{l_2}^* = \mathbb{E}_{X,Y}(m(X)-Y)^2 = \mathbb{E}_{X,Y}\left\{(\mathbb{E}(Y|X)-Y)^2\right\}$$

As explained in the introduction, there are practical situations in which the L_2 risk is not a good way of measuring the closeness of $g(X)$ to Y. We focus in this paper on the case of the Mean Absolute Percentage Error (MAPE) as an alternative to the MSE. Let us recall that the loss function associated to the MAPE is given by

$$l_{MAPE}(p,y) = \frac{|p-y|}{|y|}, \quad (5)$$

with the conventions that for all $a \neq 0$, $\frac{a}{0} = \infty$ and that $\frac{0}{0} = 1$. Then the MAPE-risk of model g is

$$L_{MAPE}(g) = L_{l_{MAPE}}(g) = \mathbb{E}\left(\frac{|g(X)-Y|}{|Y|}\right). \quad (6)$$

Notice that according to Fubini's theorem, $L_{MAPE}(g) < \infty$ implies in particular that $\mathbb{E}(|g(X)|) < \infty$ and thus that interesting models belong to $L^1(\mathbb{P}_X)$, where $\mathbb{P}_X$ is the probability measure on $\mathbb{R}^d$ induced by X.

We will also use in this paper the mean absolute error (MAE). It is based on the absolute error loss, $l_{MAE} = l_1$ defined by $l_{MAE}(p,y) = |p-y|$. As other risks, the MAE-risk is given by

$$L_{MAE}(g) = L_{l_{MAE}}(g) = \mathbb{E}(|g(X)-Y|). \quad (7)$$

3. Existence of the MAPE-regression function

A natural theoretical question associated to the MAPE is whether an optimal model exists. More precisely, is there a function m_{MAPE} such that for all models g, $L_{MAPE}(g) \geq L_{MAPE}(m_{MAPE})$?

Obviously, we have

$$L_{MAPE}(g) = \mathbb{E}_{X,Y}\left\{\mathbb{E}\left(\frac{|g(X)-Y|}{|Y|}\middle| X\right)\right\}.$$

A natural strategy to study the existence of m_{MAPE} is therefore to consider a point-wise approximation, i.e. to minimize the conditional expectation introduced above for each value of x. In other words, we want to solve, if possible, the optimization problem

$$m_{MAPE}(x) = \arg\min_{m \in \mathbb{R}} \mathbb{E}\left(\frac{|m-Y|}{|Y|}\middle| X=x\right), \quad (8)$$

for all values of x.

We show in the rest of this section that this problem can be solved. We first introduce necessary and sufficient conditions for the problem to involve finite values, then we show that under those conditions, it has at least one global solution for each x and finally we introduce a simple rule to select one of the solutions.

3.1. Finite values for the point-wise problem

To simplify the analysis, let us introduce a real valued random variable T and study the optimization problem

$$\min_{m \in \mathbb{R}} \mathbb{E}\left(\frac{|m-T|}{|T|}\right). \quad (9)$$

Depending on the distribution of T and of the value of m, $J(m) = \mathbb{E}\left(\frac{|m-T|}{|T|}\right)$ is not always a finite value, excepted for $m=0$. In this latter case, for any random variable T, $J(0)=1$ using the above convention.

Let us consider an example demonstrating problems that might arise for $m \neq 0$. Let T be distributed according to the uniform distribution on $[-1,1]$. Then

$$J(m) = \frac{1}{2}\int_{-1}^{1} \frac{|m-t|}{|t|}\,dt.$$

If $m \in]0,1]$, we have

$$\begin{aligned} J(m) &= \frac{1}{2}\int_{-1}^{0}\left(1-\frac{m}{t}\right)dt + \frac{1}{2}\int_{0}^{m}\left(\frac{m}{t}-1\right)dt + \frac{1}{2}\int_{m}^{1}\left(1-\frac{m}{t}\right)dt, \\ &= \underbrace{1-m-\frac{m}{2}\int_{m}^{1}\frac{1}{t}\,dt}_{\text{finite part}} + \underbrace{\frac{m}{2}\left(\int_{0}^{m}\frac{1}{t}\,dt - \int_{-1}^{0}\frac{1}{t}\,dt\right)}_{+\infty}, \\ &= +\infty. \end{aligned}$$

This example shows that when T is likely to take values close to 0, then $J(m) = \infty$ whenever $m \neq 0$. Intuitively, the only situation that leads to finite values is when $\frac{1}{|T|}$ as a finite expectation, that is when the probability that $|T|$ is smaller than ϵ decreases sufficiently quickly when ϵ goes to zero.

More formally, we have the following proposition.

Proposition 1. *$J(m) < \infty$ for all m if and only if*

1. $\mathbb{P}(T=0) = 0$,
2. *and*

$$\sum_{k=1}^{\infty} k\mathbb{P}\left(T \in \left]\frac{1}{k+1}, \frac{1}{k}\right]\right) < \infty, \quad \sum_{k=1}^{\infty} k\mathbb{P}\left(T \in \left[-\frac{1}{k}, -\frac{1}{k+1}\right[\right) < \infty. \quad (10)$$

If any of those conditions is not fulfilled, then $J(m) = \infty$ for all $m \neq 0$.

Proof. We have

$$J(m) = \mathbb{E}\left(\mathbb{I}_{T=0}\frac{|m-T|}{|T|}\right) + \mathbb{E}\left(\mathbb{I}_{T>0}\frac{|m-T|}{|T|}\right) + \mathbb{E}\left(\mathbb{I}_{T<0}\frac{|m-T|}{|T|}\right).$$

If $\mathbb{P}(T=0) > 0$ then for all $m \neq 0$, $J(m) = \infty$. Let us therefore consider the case $\mathbb{P}(T=0) = 0$. We assume that $m > 0$, the case $m < 0$ is completely identical. We have

$$J(m) = \mathbb{E}\left(\mathbb{I}_{T>0}\frac{|m-T|}{|T|}\right) + \mathbb{E}\left(\mathbb{I}_{T<0}\frac{|m-T|}{|T|}\right),$$